TopicPie: An Interactive Visualization for LDA-based Topic Analysis

Yi Yang, Jian Wang*, Weixing Huang, Guigang Zhang
Institute of Automation
Chinese Academy of Sciences
Beijing, China
{yangyi, guigang.zhang}@ia.ac.cn, hwx0904@vip.sina.com, jian.wang@ia.ac.cn (Corresponding author)

Abstract—**LDA-based topic analysis is widely used in text mining field. Considering the large scale of web documents, document clusters are usually analyzed instead of single ones. However, the existing visualizations of LDA-based clustering do not intuitively present contents of hot topics while maintaining the relationships between the topics and the document clusters. In this paper, we propose an integrated interactive visualization method that provides intuitive and effective views for topic popularity, topic contents, document clusters, and relationships between topics and document clusters. In this way, users can quickly identify the topic-based patterns. We show an experimental evaluation by comparing the tabular representation and our visualization. The results show that our method can significantly facilitate the topic analysis, particularly in the field of Chinese culture study.**

Keywords-Latent Dirichlet Allocation; LDA; topic model; text mining; clustering; visualization; visual analytics.

I. Introduction

Nowadays, a huge number of documents exist on the Web. Topic analysis plays an important role in web text mining in order to find the valuable topics. Latent Dirichlet Allocation (LDA) is a well-known topic analysis model that can discover the hidden topics. By integrating LDA and text clustering, web documents can be grouped regarding topics. In this way, users can quickly identify the important patterns of web documents and hot topics. In research field of Chinese culture, LDA is an important way to analyze the web Chinese cultural information. For pattern analysis of large-scale web documents, the group of documents is usually concentrated rather than single documents. The analysis focuses on the following questions: (a) *what are the hot topics?* (b) *what document clusters are there?* (c) *what document clusters contain a specified topic?* (d) *what topics are contained by a specified document cluster?* To facilitate the understanding of the results of large-scale LDA analysis and the relevant clustering, a suitable graphical representation is needed. The current visualization methods enhance the topic analysis in some specific aspects but these methods cannot quickly answer above questions. In this paper, we propose a visualization method for large-scale LDA-based topic analysis. The method presents an intuitive overview of the analysis results and dynamic detailed information by visual properties and interactions to help users easily find valuable patterns.

The rest of the paper is organized as follows. Section 2 presents the related work of LDA analysis and visualizations of LDA. Section 3 introduces the LDA-based analysis including LDA topic model and LDA-based clustering. Section 4 shows our visualization approach and the associated interactions. Section 5 gives scenarios to present the use of our visualization. Section 6 shows an experiment for evaluating the visualization. Finally, conclusion and future work are given in Section 7.

II. Related Work

Topic analysis is an essential part of text analysis. The bag-of-words methods are widely used for massive web documents mining. Vector Space Model (VSM) [1] is the basic method. TF-IDF [2], LSI [3], and pLSI [4] provide advanced concepts based on VSM. Blei et al. [5] proposed the LDA method applying Dirichlet distribution [6] in order to model the hidden topics of documents. The LDA-based methods are now the mainstream for topic modeling and extracting in text mining. Hierarchical LDA (hLDA) was described in [7]. Yao et al. [8] proposed SparseLDA to estimate Gibbs sampling distributions for accelerating LDA analysis. Researchers also put efforts in the visualization for LDA analysis. Cao et al. [9] proposed an integrated visual analytic method based on pie chart and word cloud for exploring topics in multi-relational data. Nakazawa et al. [10] visualized the LDA-based citation networks of papers by using the bundle node-link layout. Alexander et al. [11] proposed a re-orderable matrix layout for LDA in order to help find high-level patterns. Lohmann et al. [12] proposed a layered circular word cloud layout by which words of different documents can be arranged together for building an integrated visualization. De Hollander et al. [13] applied word clouds to visualize the topics of meetings. Alexander et al. [14] applied Buddy Plots to present document relationship across topic models for comparison of topics.

III. Topic Analysis

A. LDA Topic Model

The core of LDA is the Dirichlet distribution that indicates a distribution over distributions. A corpus is a document set containing with total M documents. A document is a set of N words. The corpus contains K topics. LDA is a generative process that a document d has a probability to choose a topic z and a topic has a probability to select a word w. LDA assumes Dirichlet distributions φ and θ in the selections: θ with parameter α for word selection; φ with parameter β for topic selection. With LDA, the generation process of a document in a corpus is described in [5] as follows:

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.25

1. Choose $\theta(\alpha)$ ~ Dirichlet distribution

2. For selection of each the N words:

(a) Choose a topic z_n ~ Multinomial (θ)

(b) Choose a word $w_i \sim p(w_i|z_i,\varphi(\beta))$, a multinomial probability conditioned on the topic z_i.

We use perplexity [15] to determine the optimal number of topics for LDA model and apply Gibbs sampling [16] that can quickly estimate the parameters of LDA. The results consist of two sets of vectors:

- topic-word vectors: each topic z has a probabilistic vector of words $w_1, \ldots, w_m$: $\varphi_z = \{P(w_1), P(w_2), \ldots, P(w_m)\}$
- document-topic vectors: each doc d has a probabilistic vector of topics $z_1, \ldots, z_k$: $\theta_d = \{P(z_1), P(z_2), \ldots, P(z_k)\}$

The popularity of topic z_i is defined as $\sum d_{z_i} / \sum d$, where d denotes a document and d_{z_i} is the document containing topic z_i.

B. Document Clustering Analysis

For web document clustering, we plan to build a clustering algorithm set in our system. The candidate algorithms can be selected case by case. We apply K-means [17] in this paper as an example. K-means partitions a dataset into several clusters depending on the similarity of data. The value K denoting the number of clusters needs to be given by users. The document-topic vectors of LDA analysis can be used to calculate the similarity with the K-L distance [18] for K-means clustering.

IV. TopicPie Visualization

In this paper, we propose an interactive visualization named *"TopicPie"* that aims at facilitating the LDA-based analysis by graphical representations of the large-scale multi-dimensional data of the topic analysis.

A. Requirement

TopicPie works on intuitively presenting the answers of the questions listed in Section 1 by fulfilling the requirements:

1) *Visualizing topic popularity:* to answer question (a)

2) *Visualizing topic contents:* to answer question (a)

3) *Visualizing document clusters:* to answer question (b)

4) *Visualizing topics contained by the specified document clusters:* to answer question (c)

5) *Visualizing topic distribution over document clusters:* to answer question (d)

We map the data of the LDA-based topic analysis to the comprehensible visual metaphors. In terms of the visualization requirements, the following data need to be visualized:

- Document-topic vectors: for req. (1), (4), and (5)
- Topic-word vectors: for req. (2)
- Document clusters: for req. (3)

B. Visualization

TopicPie has three basic parts (see Figure 1): topic chart view (outer donut), slice word cloud (the part attached to the topic chart view), and document cluster view (central area).

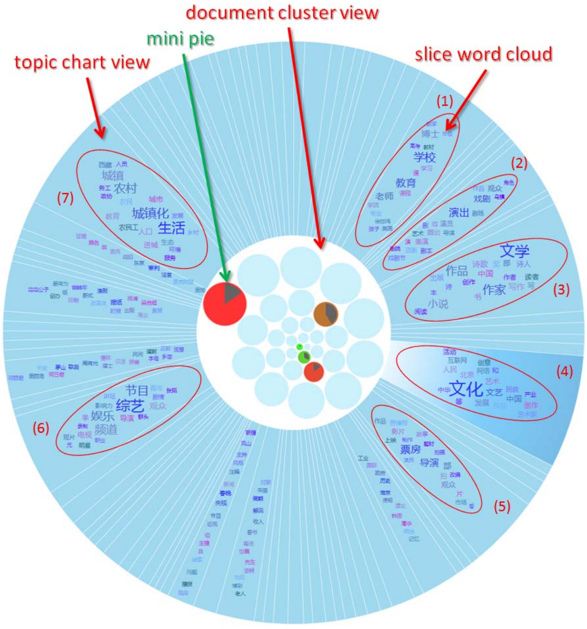


Figure 1 TopicPie Overview and hot topics (marked by red ellipses and numbers). The major words of the hot topics are described in English as follows: (1) "school, education, doctor, teacher, course, study, major, child" ; (2) "show, drama, actor, audience, play, theatre, director, role, script"; (3) literature，writer，work, novel, poetry, compose, publication, China, author, reading; (4) culture, art, Bejing, Internet, artist, development, creativity, activity, network; (5) box office, director, audience, film, work, topic, story, market, adaption; (6) entertainment, program, variety show, channel, audience, TV, director, story, pop star, recording; (7) life, country, urbanization, urban, city, population, education, Tibet, labor, ecology,urbanizing.

1) Topic Chart View

Hot topic analysis is an essential task of the topic analysis. For comparing topics, we apply a pie chart as the basic visual structure. Pie chart is a well-known statistical visualization that provides a comparable circular layout. A pie chart is divided into slices according to the numerical proportion. For TopicPie, each slice represents a topic and the slice angle is proportional to the popularity of the topic. The order of the slices depends on the order of the topic extraction order of LDA analysis.

2) Slice Word Cloud

The contents of a topic consist of a group of keywords. To intuitively visualize the topic contents, we apply a Tag Cloud-based method *"slice word cloud"* containing words of different sizes. The font size of a word is proportional to its value in the topic-word vector. The large word is more able to represent the topic than small ones. We place the slice word cloud instance into the corresponding slice. When viewing the topic chart, we can conveniently find the contents of the main topics. In order to effectively use the limited slice space, we shape the contour of a slice word cloud instance according to the corresponding slice shape. We keep words in the horizontal direction for the comfortable read style. We build an integrated view to provide the popularity of topics and the contents of topics by combining the topic chart with the slice word cloud. The major topics can be visually identified and easily understood. If a slice area is too small to hold the related slice word cloud instance, we will keep the slice area empty rather than showing words inside. Meanwhile, users can search and select the desired topic related to the small slice in a tabular view of the system GUI. In this way, the visualization requirement (1) and (2) can be satisfied, namely, the question (1) can be addressed by our visualization.

3) Document Cluster View

The central area of TopicPie is the document cluster view. In this view, a document cluster is represented as a cluster bubble. The radius of a cluster bubble indicates the number of documents in the corresponding cluster. The larger the size of a cluster is, the more the cluster is focused on. In order to facilitate the identification of the major clusters, we arrange the cluster bubbles in a descending spiral order. In this way, the document cluster view fulfills the requirement (3).

C. Interaction

In order to clearly present multi-dimensional data, we apply interactions to TopicPie. The interactions can help dynamically show the relationships between topics and document clusters from the point of views of either topic or document cluster.

1) Selection of Document Cluster

A document cluster usually contains more than one topic. Based on the topic-word vector, proportion of each topic in a document cluster can be calculated so that the contents of the document cluster are represented by quantitative combination of topics. A large-sized document cluster means that the cluster has popular contents. We provide an interaction for analyzing the topic combination. When selecting a cluster bubble, the bubble will be highlighted (see Figure 2). Meanwhile, the topic chart will be re-partitioned based on the updated proportions of topics of the selected cluster. In this way, users can intuitively carry out the quantitative analysis of topics for a cluster. In order to reduce visual collisions introduced by the topic chart update, we add animations for smoothly transforming. By the selection interaction, we dynamically present the relationships between a document cluster and topics from the point of views of documents for fulfilling the visualization requirement (4).

2) Selection of Topic Slice

When a hot topic has been identified, users are usually interested in the document clusters that contain the hot topic. Therefore, we provide a selection interaction for the topic slice. When selecting a topic slice, the slice will be highlighted by gradient color (see Figure 1 area (4)). Users can perform the following analyses by newly generated visual properties.

The first analysis is: *find the document clusters that contain the selected topic*. We use colors to encode the frequency of the topic in a document cluster. We normalize the frequencies of the selected topic for all document clusters, and then map the normalized values to rainbow color scaling from green to red, i.e., a red cluster bubble indicates that the cluster contributes a large part of the topic frequency. In this way, users can focus on this cluster in order to figure out why the topic is hot. Oppositely, a cluster bubble with green color means that the topic is only mentioned a few times in the cluster so that users do not need to focus on it. The colors of cluster bubbles are irrelevant to the colors of the slice word cloud. By using the color encoding, users can easily discover distribution patterns of a specified topic and the relevant document clusters.

The second analysis is: *determine the significance of the selected topic to a document cluster*. A document cluster may include more than one topic. The proportion of the selected topic to overall topics of the cluster is also a valuable factor. The document clusters that are dominated by the selected topic are more worth analyzing than others. In order to visualize the proportion of the selected topic in a cluster, we apply a small pie chart, namely "mini pie" on each cluster bubble (see Figure 1). The gray sector of a mini pie depends on the proportion above computed. By mini pie, users can conveniently view the proportion of the selected topic in a document. Users can analyze the topic patterns by considering both colors and the mini pie. That is, a cluster bubble that has red color and contains a mini pie with a large slice is worthy of analysis.

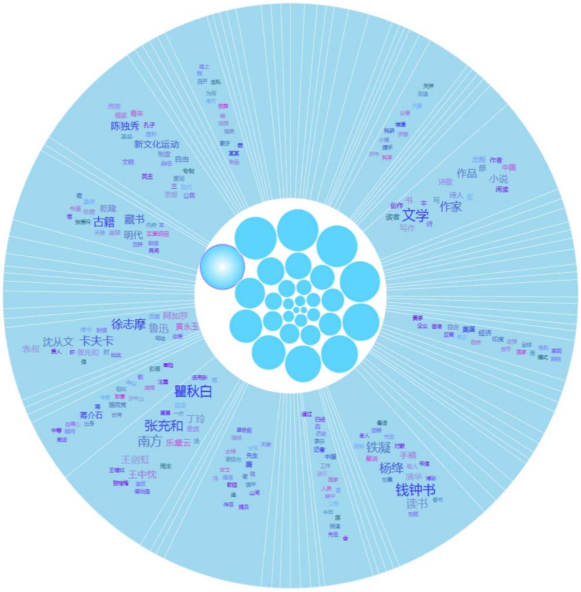


Figure 2 Interaction: select a document cluster

We apply tooltips for slices that show the words contained by the current topic in descending order. The interactions of topic slice selection can satisfy the requirement (5).

V. Case Study

We demonstrate application scenarios in this section to show the use of TopicPie. We collected over 100 thousands documents from the culture channels of hot Chinese websites.

The first scenario is to identify the hot topics. Figure 1 shows an overview of topics. We can quickly find 7 hot topics (see Figure 1) according to the slice size. By large words of the slice word cloud instances, we find that the hot topics are related to the fields of cultural creativity, literature, drama, movie, school education, variety show, and urbanization.

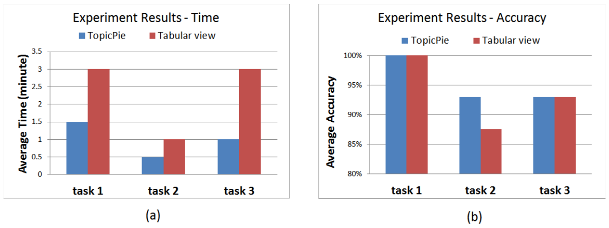
The second scenario focuses on analyzing the documents related to the most popular topic: cultural creativity. We select the relevant slice, and then five cluster bubbles are highlighted with colors and show mini pies (see Figure 1). The bubbles represent the clusters containing the selected topic. The two red bubbles mean that the selected topic is mostly contained in the clusters, while the brown bubble has a large proportion of the related documents by the mini pie. Therefore, we focus on these three clusters. When clicking the mini pie of the biggest red bubble, we can view details of the documents in the cluster.

In scenario 3, we analyze what an interesting cluster talks about. We select the biggest cluster bubble, and the topic chart view is transformed from Figure 1 to Figure 2. On the new view, the dominating slices show famous Chinese litterateurs and keywords of Chinese culture. The results show that the hot topic is about Chinese culture and litterateurs. In this case, our approach can dynamically show contents of the selected cluster, meanwhile keeping the sorting and overview of the clusters.

VI. Experimental Evaluation

We conducted an experiment to evaluate the effectiveness of TopicPie. A large-scale document set is difficult to retrieval. In order to ensure the experiment can be done in reasonable time, we applied a corpus of 500 documents regarding the Chinese culture. We set 3 tasks: Task 1: Find hot topics and describe the contents; Task 2: Find the document clusters containing the hot topics; Task 3: Find the major topics contained by a specified document cluster. There were 16 participants from the institute and University. We divided them into two groups equally. One group applied TopicPie for the tasks and another group used tabular views.

Figure 3 (a) illustrates the average time consume for the tasks. The results show that the TopicPie group completed the tasks much faster than the tabular view group. ANOVA F-test shows that for all three tasks the differences of time consume have statistical significance. Figure 3 (b) illustrates the average accuracy of the groups. The results show that the both groups have high quality achievement and there are no significant differences. We conclude that with TopicPie users significantly faster carry out the LDA-base topic analysis than using the tabular views while obtaining the similar accuracies.



(a)

(b)

Figure 3 Experiment result: (a) average time and (b) average accuracy

VII. Conclusion

In this paper, we have introduced an integrated interactive visualization *TopicPie*. It aims at facilitating the LDA-based topic analysis for massive web documents. Using TopicPie, users can intuitively identify hot topics and the relationships between topics and document clusters. We have demonstrated application scenarios regarding the topic analysis for the Chinese culture study as well. With an experimental evaluation, we concluded that TopicPie can significantly outperform the traditional tabular representations. In the future, we will carry out experiments to compare the existing visualizations of LDA with our visualization approach. Moreover, we will focus on the visualizations for the hierarchical LDA-based topic analysis.

Acknowledgment

We would like to thank all colleagues and students who helped for our work and thank the following support: (1) Research of Big Data Collection and Analysis for Public Culture Services (Project of National System Design of Public Cultural Service System 2015/2016) (2) Support Program of the National '12th Five-Year-Plan of China' under Grant No. 2015BAK25B04 and Grant No. 2015BAK25B03; (3) National Cultural Resource Sharing Project (key project of '13th Five-Year Plan of China'); (4) China Association for Science and Technology.

References

[1] G. Salton, A. Wong , C. S. Yang, A vector space model for automatic indexing, Communications of the ACM, v.18 n.11, p.613-620, 1975.

[2] G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[3] Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, 1988, pp. 36–40.

[4] T. Hofmann. Probabilistic latent semantic analysis. In Proc. of Uncertainty in Artificial Intelligence, UAI'99. Stockholm, 1999.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

[6] Bela A. Frigyik, Amol Kapila, and Maya R. Gupta. Introduction to the Dirichlet Distribution and Related Processes. ee.washington.edu. Retrieved 14 May 2015.

[7] D. Blei, T. Griffiths, and M. Jordan. The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57(2):1–30, 2010.

[8] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In KDD, 2009.

[9] Nan Cao, David Gotz, Jimeng Sun, Yu-Ru Lin and Huamin Qu. SolarMap: Multifaceted Visual Analytics for Topic Exploration. Data Mining (ICDM), 2011 IEEE 11th International Conference on, 11-14 Dec. 2011, Vancouver, BC, Canada, 101 - 110.

[10] Nakazawa, R.; Itoh, T.; Saito, T. Information Visualisation (iV), 2015 19th International Conference on 22-24 July 2015, 283-289, Barcelona.

[11] Eric Alexander, Joe Kohlmann, Michael Witmore, Robin Valenza, Michael Gleicher. Serendip: Topic Model-Driven Visual Exploration of Text Corpora. Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on 25-31 Oct. 2014,Paris, 173 - 182.

[12] Lohmann, S.; Heimerl, F. ; Bopp, F.; Burch, M. et al. ConcentriCloud: Word Cloud Visualization for Multiple Text Documents.IEEE Information Visualisation (iV), 2015 19th International Conference on 22-24. July 2015,114 - 120,Barcelona, 2015.

[13] De Hollander, G.; Marx, M. Summarization of meetings using word clouds.Computer Science and Software Engineering (CSSE), 2011 CSI International Symposium on 15-16 June 2011, Tehran, 54-61.

[14] Alexander, E.; Gleicher, M.; Task-Driven Comparison of Topic Models. Visualization and Computer Graphics, IEEE Transactions on 13 August 2015, Volume:22,Issue:1,320 - 329.

[15] HM Wallach, I Murray, R Salakhutdinov, and DM Mimno. Evaluation methods for topic models. In Proc. of ICML 2009, page 139.

[16] T. Griffiths and M. Steyvers. Finding scientiffic topics. In PNAS, volume 101, pages 5228-5235, 2004.

[17] Kotsiantis, S. and Pintelas, P. E. 2004. Recent advances in clustering: A brief survey. WSEAS. Trans. Inform. Sci. Appl. 1, 1, 73–81, 2004.

[18] Kullback, S. Letter to the Editor: The Kullback-Leibler distance. The American Statistician 41 (4): 340-341, 1987.